The Annals of Applied Statistics
2011, Vol. 5, No. 2B, 1360–1378
DOI: 10.1214/10-AOAS438

FALSE DISCOVERY RATES IN SOMATIC MUTATION STUDIES OF CANCER[1]

By Lorenzo Trippa and Giovanni Parmigiani

Dana-Farber Cancer Institute and Harvard School of Public Health

The purpose of cancer genome sequencing studies is to determine the nature and types of alterations present in a typical cancer and to discover genes mutated at high frequencies. In this article we discuss statistical methods for the analysis of somatic mutation frequency data generated in these studies. We place special emphasis on a two-stage study design introduced by Sjöblom et al. [*Science* **314** (2006) 268–274]. In this context, we describe and compare statistical methods for constructing scores that can be used to prioritize candidate genes for further investigation and to assess the statistical significance of the candidates thus identified. Controversy has surrounded the reliability of the false discovery rates estimates provided by the approximations used in early cancer genome studies. To address these, we develop a semiparametric Bayesian model that provides an accurate fit to the data. We use this model to generate a large collection of realistic scenarios, and evaluate alternative approaches on this collection. Our assessment is impartial in that the model used for generating data is not used by any of the approaches compared. And is objective, in that the scenarios are generated by a model that fits data. Our results quantify the conservative control of the false discovery rate with the Benjamini and Hockberg method compared to the empirical Bayes approach and the multiple testing method proposed in Storey [*J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** (2002) 479–498]. Simulation results also show a negligible departure from the target false discovery rate for the methodology used in Sjöblom et al. [*Science* **314** (2006) 268–274].

1. Introduction. The systematic investigation of the genomes of human cancers has recently become possible with improvements in sequencing and bioinformatic technologies. Sjöblom et al. (2006) and Wood et al. (2007) determined the sequence of comprehensive collections of coding genes (CCDS

Received May 2010; revised October 2010.
[1]Supported in part by NSF Grant DMS-03-42111.
Key words and phrases. Cancer genome studies, genome-wide studies, false discovery rate, multiple hypothesis testing, somatic mutations.

and RefSeq) in colorectal and breast cancers, and provided a catalogue of somatic mutations. In this context, a somatic mutation is a tumor-specific mutation not present in the germline of the patient whose tumor contained it. Subsequently, Greenman et al. (2006) investigated somatic mutations in the coding exons of 518 protein kinase genes in a large and diverse set of human cancers. More recently, mutation data from glioblastoma tissues have been studied in the The Cancer Genome Atlas project (2008) and Parsons et al. (2008), and from pancreatic cancer in Jones et al. (2008). Statistical analysis of the data generated in these studies poses new challenges that are worthy of careful consideration. Greenman et al. (2006) have provided an in-depth analysis of data generated by one-stage studies. To make optimal use of sequencing resources, Sjöblom et al. (2006) introduced a two-stage design, with the stages termed "Discovery" and "Validation." The Discovery Stage consists of a catalog of mutations in all genes considered, for example, all genes in the CCDS database. This design permitted selection of the subset of genes that harbored at least one somatic mutation, termed "Discovered." This subset was further investigated in a Validation Stage which cataloged somatic mutations in discovered genes in an independent set of tumor samples. Genes that were mutated in at least one tumor in the Validation set were termed "Validated." In this article we consider this two-stage design. Sjöblom et al. (2006) and Wood et al. (2007), adopting this experimental design, discovered that among the genes whose mutations are likely responsible for carcinogenesis, the majority, the "hills," had mutations in small subgroups of cases, while only a handful of genes, the "mountains," were mutated in large subgroups. Thus, the hills and not the mountains dominate the cancer genome landscape. This imbalance emphasizes the importance of statistical methods in identifying the mutations involved in the carcinogenesis process.

The somatic mutations found in cancer tissues are either "drivers" or "passengers" [Wood et al. (2007)]. Driver mutations are causally involved in the neoplastic process and are positively selected during tumorigenesis. Passenger mutations provide no positive or negative selective advantage to the tumor but are retained by chance during repeated rounds of cell division and clonal expansion. The overarching goal of the statistical analysis of cancer mutation data is to identify genes that are most likely to contain driver mutations on the basis of their mutation type and frequency. This is done by quantifying the evidence that the mutations in a gene reflect underlying mutation rates that are higher than the passenger rates.

Early cancer genome projects provided a rank order of genes by their potential to be drivers of carcinogenesis based on mutation frequencies in tumors as well as the genes' size and nucleotide compositions. To provide an indication of the significance of lists of possible driver genes, they also provided estimates of the false discovery rate (FDR), that is, the expected proportion of putative drivers that are actually passengers, or the proportion of

erroneously rejected null hypotheses [Benjamini and Hochberg (1995)]. Since the seminal article of Benjamini and Hochberg (1995), several authors have proposed alternative methods that control the FDR with improved operating characteristics. Important contributions include relaxing the independence assumption among the test statistics and handling discrete statistics [Benjamini and Yekutieli (2001)]. Reviews are given in Dudoit, Shaffer and Boldrick (2003), Cheng and Pounds (2007) and Farcomeni (2008). Despite a rich literature, applications often include heuristic arguments and, in most cases, the choice of a method is far from trivial: a trade-off arises between methods which have been proved through rigorous analytical arguments to control the FDR under a specified threshold and less conservative procedures that rely on heuristic arguments, examples and asymptotic theory. The trade-off between methodological rigor and operating characteristics becomes particularly relevant, for example, when the independence hypothesis is inappropriate or when data are highly discrete. Applications often ignore these problems.

In this article we consider methods that are specific to somatic mutation analysis, and assess them by a novel model-based approach. Our idea is to develop a "super partes" model that can be used to provide highly realistic artificial datasets for method evaluation. We refer to these as data-driven simulated scenarios. Our scheme for evaluating alternative methods can take into account possible discrepancies between methods' assumptions and the data. We apply this comparison method by revisiting the Sjöblom et al. (2006) methodology as well as the alternative approaches proposed shortly afterward by several groups [Forrest and Cavet (2007), Getz et al. (2007), Rubin and Green (2007) and Parmigiani et al. (2007a, 2007b)]. We discuss the application of this scheme to the data collected in Wood et al. (2007).

The article is structured in 5 sections. In Section 2 we introduce the notation for modeling mutation counts in tumor tissues and present our probability model. In Section 3 we briefly review techniques for controlling or estimating false discovery rates, and specific approaches for the analysis of somatic mutation data. In Section 4 we compare these approaches. Final remarks are given in Section 5.

2. Data-driven simulation scenarios.

2.1. *Approach.* The idea of data-driven simulations is structured in two steps. First we select a flexible probability model for the observed mutations. The model includes model-specific parameters, genes-specific latent variables and observable mutation counts. The model is consistent with widely established assumptions on carcinogenesis. Prior distributions on the unknown model parameters are specified by using both external experimental results

and heuristic data-driven approaches. Second, we infer the genes-specific latent variables, including driver status, through a Monte Carlo Markov chain (MCMC) algorithm and generate simulated data sets consistent with the Wood et al. (2007) study.

Throughout the article this Bayesian model is used only for generating simulated data sets that are highly consistent with the observed data in Wood et al. (2007) and are impartial to the FDR approaches examined. In principle, the model could be used directly for selecting genes having high posterior probabilities of being drivers. Other potential applications of the model include (i) predicting the experimental outcomes for ongoing studies, (ii) optimally choosing the resource allocation for the validation stage using a decision theoretic approach and, more generally, (iii) developing adaptive strategies for sequencing experiments. Nevertheless, here we prefer limiting our comparisons to highly computationally efficient methods whose operating characteristics can be assessed via Monte Carlo techniques over a large number of data sets. Moreover, it would be circular to simultaneously use our Bayesian model for FDR estimation and for generating data for evaluation. As noted in Getz et al. (2007), the performance of some methods could be sensitive to the distribution underlying the experimental data; using a probability model that fits the observed data, but also averages across plausible values of the unknown parameters, allows us to provide appropriately objective comparisons.

2.2. *Sampling model.* While cancer genome sequencing projects produce a wealth of information, in this article we will focus on the somatic mutation counts, broken down by gene and context, and considered separately for the Discovery and Validation Stages. Table 1 summarizes the notation we will use. Our model applies to a single disease, say, colorectal cancer, at a time.

TABLE 1
Summary of notation for the data produced by the study for the gth gene, and associated gene-specific parameters

Mutation counts	
X^1_{gm}	number of mutations of type m detected in gene g in the Discovery Stage.
X^2_{gm}	number of mutations of type m detected in gene g in the Validation Stage.
Coverage	
T^1_{gm}	coverage of type m in gene g in the Discovery Stage.
T^2_{gm}	coverage of type m in gene g in the Validation Stage.
Mutation rates	
γ^1_m	rate of mutation of type m in the Discovery Stage.
γ^2_m	rate of mutation of type m in the Validation Stage.
θ_g	multiplicative gene-specific random effect.

The probability model is defined on the basis of a few well-established assumptions that allow to specify a distribution of mutation counts conditional on unknown gene-specific mutation rates. Using latent variables, we specify a model allowing for unknown composition of the genome in terms of passengers and drivers, and for heterogeneous mutation rates across driver genes. More formally, we assume that, for each gene and sample, the number of mutations of type m, that is, the number of identical mutations that can occur only in a specific context, for example, C to G in a CpG locus, is a mixture of Poisson distributions,

$$X_{igm}|\theta_g, \rho_i, \eta_m \sim \text{Poisson}(\rho_i \eta_m \theta_g T_{igm}) \quad \text{and} \quad \theta_g \overset{\text{i.i.d.}}{\sim} F, \tag{2.1}$$

where i indexes a tumor, g indexes a gene and T_{igm} denotes the coverage, or the number of successfully sequenced nucleotides susceptible to mutation of type m in gene g and sample i. The parameter η_m represents the rate of mutations of type m among passengers. The multiplicative factor ρ_i is designed to capture the fact that the abundance of mutations varies across tumors. The transition of a tissue from normal to cancer can be described as a progressive accumulation of mutations, some of them are drivers and some are passengers. This dynamic varies across tumors; such heterogeneity is the focus of a significant portion of cancer research; see Stratton, Campbell and Futreal (2009) for a stimulating discussion. Our model, as well as those used in the previously mentioned cancer studies, describes a snapshot of this dynamic process at the time of sequencing. The product $(\rho_i \eta_m)$ can be interpreted as the rate of mutations of type m in tumor i, assuming that the nucleotide is part of a passenger gene. The gene-specific latent variables θ_g capture gene-specific variation across the genome; if $\theta_g = 1$, gene g is a passenger, while higher values identify the drivers. Our model assumes that the rates of mutation across different types m of mutations in a single gene g are proportional to the rates of mutation in a passenger gene. Finally, F is the distribution of θ_g's across the genome. It allows for values of 1 or bigger and allows for a concentration of mass on the value of 1, corresponding to the passengers.

The overall structure of the model reflects the assumption that the drivers have higher rates of mutation than the passengers. The analyses in Sjöblom et al. (2006) and Wood et al. (2007) were aimed at selecting cancer genes with mutation rates higher than the hypothesized passenger rates. The use of the Poisson distribution in (2.1) is motivated by the fact that, under mild assumptions, it well approximates a more rigorous multinomial model [Wood et al. (2007)] obtained by modeling possible mutations in a single gene as binary variables.

Values of individual passenger rates $(\rho_i \eta_m)$ can be obtained using data external to the somatic mutations counts [Sjöblom et al. (2006), Wood et al.

(2007)]. This allows us to simplify the model (2.1) by collapsing data across patients. Once the intensities $(\rho_i \eta_m)$ are known, the collapsed counts data $(X_{gm} = \sum_i X_{igm})$ are sufficient statistics for evaluating the likelihood function. This allows for a considerable reduction of the computational requirements. We will thus use the alternative representation of (2.1):

$$X_{gm}|\theta_g, \gamma_m \sim \text{Poisson}(\gamma_m \theta_g T_{gm}) \quad \text{and} \quad \theta_g \overset{\text{i.i.d.}}{\sim} F, \tag{2.2}$$

where X_{gm} is the total number of mutations of type m harbored in the gth gene, $T_{gm} = \sum_i T_{igm}$ and $T_{gm}\gamma_m = \sum_i \rho_i \eta_m T_{igm}$.

Multistage designs are attractive strategies for identifying cancer genes; the first stage indicates a subset of genes that are more likely to be drivers and, in the subsequent phase, only this subset is analyzed. Relevant cost–effectiveness analysis and comparisons of alternative designs for genome-wide studies are illustrated in Satagopan et al. (2002), Satagopan and Elston (2003), Satagopan, Venkatraman and Begg (2004), Kraft (2006), Skol et al. (2006), Wang and Stram (2006) and Parmigiani et al. (2009). In the studies discussed by Sjöblom et al. (2006) and Wood et al. (2007), at the end of the first stage all genes which harbored one or more mutations are considered for further study. We will use the notation X^1_{gm} and X^2_{gm} for denoting the number of mutations in the two phases. Similarly, T^1_{gm} and T^2_{gm} denote the coverages and (γ^1_m, γ^2_m) the rates for the discovery and validation phases. Model (2.2) can be adapted to the two-stage design as follows:

$$\begin{aligned} X^1_{gm}|\theta_g, \gamma^1_m &\sim \text{Poisson}(\gamma^1_m \theta_g T^1_{gm}), \\ X^2_{gm}|X^1_{gm}, \theta_g, \gamma^1_m &\sim \begin{cases} 0, & \text{if } X^1_{gm} = 0, \\ \text{Poisson}(\gamma^2_m \theta_g T^2_{gm}), & \text{if } X^1_{gm} > 0, \end{cases} \\ \theta_g &\overset{\text{i.i.d.}}{\sim} F. \end{aligned} \tag{2.3}$$

In these expressions the coverages T^1_{gm} and T^2_{gm} are considered fixed. While some variation may be experimentally observed, this is unlikely to be related to a gene's driver status, and, thus, it is appropriate to model the data conditionally on the coverages. For our purpose, the following three considerations are critical:

Two-stage design. Only genes that harbor at least one mutation in the Discovery Stage are sequenced in the Validation Stage. This screening condition needs to be taken into account when assessing significance using p-values or other methods that rely on the sampling distribution.

Coverage. The number of nucleotides successfully sequenced is generally smaller than the gene length times the number of tumors analyzed. For example, certain exons may be technically challenging to sequence. It is appropriate to apply stringent quality criteria to sequencing data, which

lead to the exclusion of nucleotides whose sequence could not be identified with certainty. Nucleotides excluded, or not covered, should not be included in statistical evaluations. In what follows p-values and other statistics used for prioritizing putative driver genes are computed taking into account, for each gene, which loci have been sequenced.

Context. The third consideration involves the observed bases in the mutations. In sequencing studies, the precise bases that comprise the mutation, as well as the neighboring bases, termed "mutation contexts," are important. We use a classification of contexts provided in Wood et al. (2007). This is a partition of the sequenced basis. For our purpose, it is relevant that each subset has specific rates of mutation under the passenger status. This fact implies that the priority given to a gene in further studies and the statistical significance of a gene should depend not only on the number of nucleotides but also on the gene-specific basis sequence.

2.3. *A Bayesian approach to generating simulation scenarios.* For our analysis we propose embedding the sampling model (2.3) in a Bayesian semiparametric model. This will allow us to generate possible scenarios consistent with the data from the Wood et al. (2007) study. We use a Dirichlet process prior [Ferguson (1973)] for the unknown distribution F,

$$F \sim \text{Dirichlet}(\mathcal{A}), \tag{2.4}$$

where $\mathcal{A}$ is a positive measure on $[1, \infty)$. Reviews on the Dirichlet process and applications in biostatistics are given in Dunson (2010) and Müller and Quintana (2004).

We chose the Dirichlet mixture model because of its flexibility compared to alternative parametric prior distributions. Simple preliminary analysis of the mutation data shows that the so-called mountains can have mutation rates over 100-folds higher than the passengers, while hills have markedly lower rates. To capture both the mountains and the hills, we specify a sufficiently flexible prior on the unknown mixing distribution F. As shown in Venturini, Dominici and Parmigiani (2008), Bayesian mixtures model effectively heavy tail distributions; in contrast, the posterior behavior of more parsimonious parametric models can be strongly biased.

In order to generate scenarios consistent with the data in Wood et al. (2007), we use a Monte Carlo Markov chain algorithm discussed in Escobar and West (1995). The sampler is based on the Polya urn representation of the Dirichlet process [Blackwell and MacQueen (1973)] and has been studied for posterior simulation under the generic Dirichlet mixture model

$$p(X_1, \ldots, X_n | F) = \prod_{i=1}^{n} \int p(X_i|\theta)\, dF(\theta), \qquad F \sim \text{Dirichlet}(\mathcal{A}).$$

The only condition for implementing the sampling scheme of Escobar and West (1995) is that for every subset $\{i_1, \ldots, i_m\}$ of distinct integers ranging from 1 to n, the integral

$$\int \prod_{j=1}^{m} p(X_{i_j}|\theta)\, d\mathcal{A}(\theta) \tag{2.5}$$

can be easily computed. To this end, we specify the prior parameter $\mathcal{A}$ proportional to a spiked distribution, including a point mass at $\{1\}$ and a shifted gamma distribution with support $[1, \infty)$, that is,

$$\mathcal{A}(dx) \propto \delta_1(dx) + cI(x \geq 1)e^{-a(x-1)}(x-1)^{b-1}\, dx, \tag{2.6}$$

where $\delta_1(\cdot)$ is a Dirac measure, $I(\cdot)$ is the indicator function and a, b, c are strictly positive. It can be verified that this choice of $\mathcal{A}$ allows us to analytically solve integral (2.5).

We chose the mean of the random distribution F by a simple procedure. We recall that the centering distribution of the Dirichlet process is $\mathcal{A}(dx)/\mathcal{A}([1, \infty))$; for every subset B of the real line, if $0 < \mathcal{A}(B) < \mathcal{A}([1, \infty))$, then $F(B)$, a priori, is beta distributed, with mean $\mathcal{A}(B)/\mathcal{A}([1, \infty))$; see Ferguson (1973). In order to specify the centering distribution, we first compute the maximum likelihood estimator $\hat{F}$ of the mixing distribution in (2.3). Then, we specify the parameter $\mathcal{A}(\{1\})/\mathcal{A}([1, \infty))$, this is the a priori expectation of the unknown proportion of passenger genes. In what follows $\mathcal{A}(\{1\})/\mathcal{A}([1, \infty))$ is set equal to $\hat{F}([1, 2))$. Finally, the parameterization of the centering distribution is completed by setting a and b in (2.6) so that the means and variances of the two distributions $\hat{F}(dx)$ and $\mathcal{A}(dx)/\mathcal{A}([1, \infty))$ are identical. The steps outlined have clear interpretations, nevertheless, the choice $\mathcal{A}(\{1\})/\mathcal{A}([1, \infty)) = \hat{F}([1, 2))$ is, to some extent, arbitrary; we also implemented posterior inference for alternative prior parameterizations: these include $\mathcal{A}(\{1\})/\mathcal{A}([1, \infty))$ equal to $\hat{F}([1, 1.5))$, $\hat{F}([1, 3))$ and $\hat{F}([1, 4))$. We did not observe marked sensitivity; for example, the ratio between the maximum and the minimum posterior estimates of the number of drivers is equal to 1.06.

We use the contexts and mutations classification discussed in Wood et al. (2007). This classification is important because it allows us to account for variations in the rates of mutations for passengers across loci, with rates depending on the basis sequences. This classification includes 25 possible types of mutations ($m = 1, \ldots, 25$). The rates for the 1st and the 2nd stage (γ_m^1 and γ_m^2) were measured using SNP data. The SNP-based approach estimates the passenger mutation rates by comparing the nonsynonymous to synonymous mutation ratios in cancer and normal tissues. This approach has been considered in Wood et al. (2007). It estimates the passenger rates using sequencing data from loci which are known for not contributing to

carcinogenesis, and thus are not positively selected during carcinogenesis; this characteristic is the defining feature of passenger genes.

A key advantage of the Bayesian estimate of the mixing distribution, compared to the maximum likelihood estimate $\hat{F}$, is that it allows us to fully take into account the uncertainty on the distribution of rates across genome, and to produce data sets under many different plausible versions of this distribution.

The MCMC algorithm outlined, after a sufficient number of burn-in iterations, produces approximate samples from the conditional distribution of the latent variables given the data. The number of burn-in iterations can be assessed by means of standard diagnostic procedures for MCMC methods; see, for example, Smith (2007). Each iteration of the MCMC algorithm provides a collection of θ_g's which is used as a simulation scenario. For each scenario we generate a single data set X using (2.3). Each scenario can be used to evaluate a given list of putative drivers by checking the proportion of genes in the list for which $\theta_g = 1$.

To highlight the excellent fit we obtain, Figure 1 provides an overview of 10,000 scenarios sampled by means of the proposed approach. Each iteration attempts to reproduce the experiment on colorectal tumors discussed in Wood et al. (2007). They considered 18,190 genes. During the discovery phase, all genes were sequenced in 11 tumors. During the validation phase,

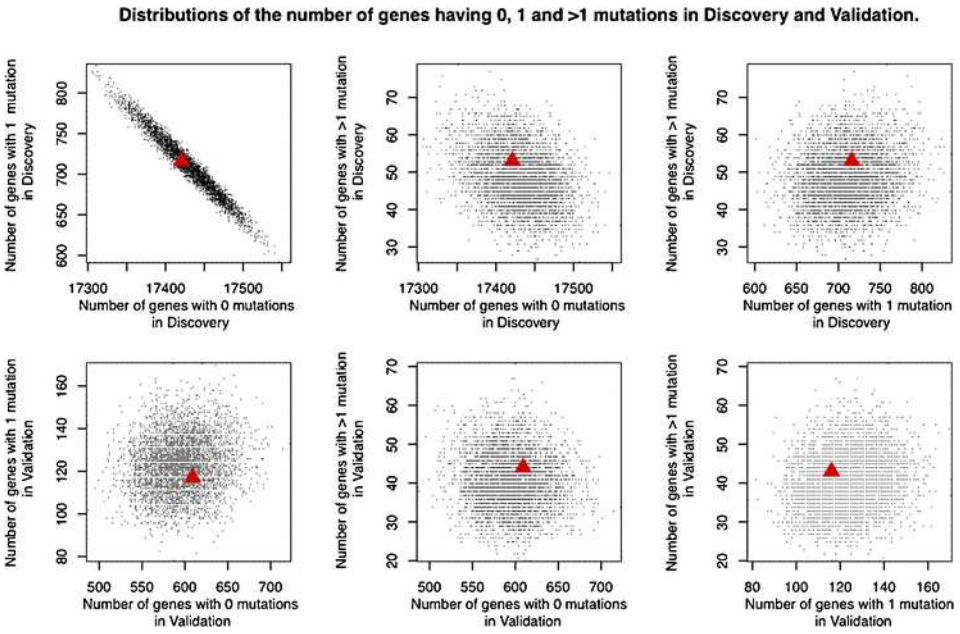

FIG. 1. *Correspondence between simulated and observed counts of mutated genes in 10,000 simulated scenarios.*

769 genes were sequenced in 24 tumors. In the discovery phase, 17,421 genes did not harbor any mutation, 716 genes harbored 1 mutation and 53 genes harbored more than 1 mutation. In the validation phase, 609 did not harbor any mutation, 115 harbored 1 and 45 harbored more than 1. Figure 1 represents the distribution, across the simulated scenarios, of the number of genes harboring 0, 1 or more than 1 mutations in the two stages. As appropriate, scenarios vary in their number of mutated genes found, while distributions are centered around the observed data.

The data-driven simulation approach requires the variability across simulations to be consistent with the data. In principle, if the experiment was repeated several times, one would like to observe a similar degree of variability across simulations and across experiments. In practice, the degree of variability across simulations can be critically evaluated by means of inferential arguments. The bootstrap method is specifically designed for predicting, on the basis of a single experiment, the degree of variability across independent replicates of the experiment. We compared the degree of variability across simulations, illustrated in Figure 1, with the variability estimates obtained by bootstrapping. Our application of the bootstrap builds on parametric estimates, for each gene, of the probabilities $p(\sum_m X^1_{gm} = 1)$, $p(\sum_m X^1_{gm} > 1)$, $p(\sum_m X^1_{gm} \geq 1, \sum_m X^2_{gm} = 1)$ and $p(\sum_m X^1_{gm} \geq 1, \sum_m X^2_{gm} > 1)$, obtained by fitting logistic binary regression models; these estimates are functions of the gene-specific coverages T^1_{gm}. The degree of uncertainty represented in Figure 1 agrees with the bootstrap estimates; the ratios between the standard deviations of the six univariate empirical distributions illustrated in Figure 1 and the corresponding bootstrap estimates range between 1.03 and 1.17. Note that the six variables considered in Figure 1 define a coarse partition of the genes; more generally, one can use the bootstrap method for assessing the degree of variability of other marginal distributions.

3. Alternative methods for controlling the FDR. In this section we review alternative FDR methods, we will then compare their operating characteristics by means of the simulations described in the previous section.

3.1. *The Benjamini and Hochberg, and the Storey procedures.* Benjamini and Hochberg (1995) considered which of the null hypotheses $(H_g, g = 1, \ldots, G)$, if any, should be rejected given p-values $(Z_g, g = 1, \ldots, G)$, one for each hypothesis. They proposed a procedure for rejecting a (possibly empty) subset of hypotheses so as to control the FDR, that is,

$$\mathbb{E}\left(\frac{\text{number of erroneously rejected hypotheses}}{\text{number of rejected hypotheses}}\right), \tag{3.1}$$

with the proviso that the above ratio is 0 when none of the hypotheses is rejected. The expectation in (3.1) is with respect to the unknown joint distribution of the test statistics Z_g. The input of the procedure is the vector

$(Z_g, g = 1, \ldots, G)$ and the output is the subset of rejected hypotheses. The p-values corresponding to the true null hypothesis are independently uniformly distributed. The FDR is an attractive error measurement in many applications with massive multiple hypotheses testing; we refer to Dudoit, Shaffer and Boldrick (2003) for a comparison with alternative error measurements.

Let $(Z_{(1)}, \ldots, Z_{(G)})$ be the sorted values in ascending order of the p-values $(Z_1, \ldots, Z_G)$ and let $\alpha \in (0,1)$ be any desired upper bound for the FDR. Benjamini and Hochberg (1995) proved that the procedure that rejects the hypotheses with a p-value lower than

$$\max\left\{\{0\} \cup \left\{Z_{(g)} : Z_{(g)} < \alpha \frac{g}{G}\right\}\right\} \tag{3.2}$$

controls the FDR below α. They show that, for every hypothetical proportion p_0 of the true null hypothesis,

$$\text{FDR} \leq p_0 \alpha. \tag{3.3}$$

The above inequality shows that the procedure is conservative. Storey (2002) studied an alternative step-up method which starts from an estimate $\hat{p}_0$ of the proportion of the true null hypothesis and then set a threshold similar to (3.2) indicating the rejection region. The estimate $\hat{p}_0$ and the inequality (3.3) are used for inflating the upper bounds $\alpha(g/G)$, $g = 1, \ldots, G$, in (3.2). The estimate of p_0 is based on the right tail of the empirical distribution of the p-values.

3.2. *The empirical Bayes method.* The use of empirical Bayes procedures for estimating the FDR has been discussed by several authors, including Efron et al. (2001), Efron (2003) and Dudoit, Gilbert and Laan (2008). In this approach, one computes a data summary, or score Z_g, that captures departure from the null hypothesis, such as a p-value, a likelihood ratio or other statistics. In our case, Z_g should capture evidence for the rate of mutation for gene g being higher than the passenger rate. The empirical Bayes method is based on a mixture representation of the distribution of these scores:

$$f(z) = p_0 f_0(z) + (1 - p_0) f_1(z); \tag{3.4}$$

$f(\cdot)$ is the distribution of a randomly selected score, p_0 is the unknown proportion of true passengers, $f_0(\cdot)$ is the distribution of a score randomly selected among passengers and, finally, $f_1(\cdot)$ is the distribution of scores among drivers. The objectives are estimating the conditional probabilities

$$\frac{(1 - p_0) f_1(Z_g)}{p_0 f_0(Z_g) + (1 - p_0) f_1(Z_g)}$$

and identifying a rejection region $\mathcal{R}$ containing the more significant scores, such that

$$\frac{\int_{\mathcal{R}} p_0 f_0(z)\,dz}{\int_{\mathcal{R}} p_0 f_0(z) + (1-p_0) f_1(z)\,dz} \leq \alpha; \tag{3.5}$$

where α is the target ratio between the mistakenly rejected hypothesis and the total number of rejections.

The distribution f_0 can be approximated simulating the scores assuming the genome only consists of passenger genes [Wood et al. (2007)], while the distribution f and the proportion p_0 are usually estimated by smoothing the scores' empirical distribution. Finally, expression (3.5) is used for rejecting a subset of null hypothesis in such a way that the estimated proportion of erroneously rejected null hypothesis is lower than α.

There is an important difference between the Benjamini and Hochberg (1995) method and the empirical Bayes method. The former rejects a subset of a list of null hypotheses and controls the expected proportion of erroneously rejected hypotheses. The latter, for a generic rejection region, estimates the proportion of true null hypotheses; the investigator can then select a subset of hypothesis such that the estimate is lower than a desired threshold α.

3.3. *The CaMP score.* Sjöblom et al. (2006) introduced the cancer mutation prevalence (CaMP) score, to provide a ranking of the Validated genes and select promising candidates. The score is based on the probability of observing the number of actually found mutations if the gene was a passenger gene, using a binomial model:

$$p_g = \begin{cases} \displaystyle\prod_m \binom{T^1_{gm}}{X^1_{gm}} (\gamma^1_m)^{X^1_{gm}} (1-\gamma^1_m)^{T^1_{gm}-X^1_{gm}}, & \displaystyle\sum_m X^1_{gm} + X^2_{gm} = 0, \\ \displaystyle\prod_{j=1}^{2} \prod_m \binom{T^j_{gm}}{X^j_{gm}} (\gamma^j_m)^{X^j_{gm}} (1-\gamma^j_m)^{T^j_{gm}-X^j_{gm}}, & \displaystyle\sum_m X^1_{gm} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

Recall that γ^1_m and γ^2_m are expected proportions of nucleotides, in a passenger gene, harboring mutations of type m. The model takes into account the two-stage experimental design.

We then rank the p_g's and call q_g the resulting ranks. The CaMP score is defined as

$$\text{CaMP}_g(X^1_{g1}, \ldots, X^1_{gM}, X^2_{g1}, \ldots, X^2_{gM}) = -\infty \qquad \text{if } \sum_m X^2_{gm} = 0$$

and

$$\text{CaMP}_g(X^1_{g1}, \ldots, X^1_{gM}, X^2_{g1}, \ldots, X^2_{gM}) = -\log_{10}(p_g/q_g) \qquad \text{if } \sum_m X^2_{gm} > 0.$$

The top row corresponds to genes that are eliminated at the Discovery or Validation Stage. The goal of the CaMP score is to rank genes according to the strength of the evidence that they may be mutated at rates higher than the passenger rates. An advantage of CaMP scores is that they can be easily computed. This definition can also be seen as an approximation of the Benjamini and Hochberg (1995) procedure. In Sjöblom et al. (2006) a threshold of 1 on the CaMP scores was considered to generate a list of putative drivers, with the goal of producing a list having approximately 10% of erroneously discovered drivers. The probabilities p_g are not p-values because they are not tail probabilities, but can be used as approximation of the p-values if the expected number of mutations in each gene is close to 0.

Forrest and Cavet (2007) proposed to use tail probabilities for controlling the FDR, though they use a sampling model that does not account for the two-stage design. Getz et al. (2007) emphasized that p-values can be used for controlling the FDR and proposed alternative test statistics. The closer test statistics to the CaMP score are obtained by computing the distribution of p_g, under the hypothesis that the gth gene is a passenger gene, and evaluating the resulting tail probability. This procedure produces p-values which can be used for controlling the FDR by applying the Benjamini and Hochberg (1995) method. Getz et al. (2007) noted that the CaMP score is not a monotone transformation of the probabilities p_g; that is, given two genes g' and g'', it can happen that $p_{g'} < p_{g''}$ and $\text{CaMP}_{g''} < \text{CaMP}_{g'}$. Getz et al. (2007) also discussed the idea of controlling the FDR by using the log-likelihood ratio, that is,

$$\log(p(X_g|\theta_g)) - \log(p(X_g|\hat{\theta}_g)),$$

where θ_g represents the hypothesis that gene g is a passenger and $\hat{\theta}_g$ is the gene-specific maximum likelihood estimate. Rubin and Green (2007) proposed to use the tail probabilities $p(\sum_{m=1} X_{gm} \geq x)$, based on the aggregate number of mutations in a gene. These critiques of the analysis in Sjöblom et al. (2006) and the alternatives proposed by these authors suggest the idea of systematically assessing the operative characteristics of alternative approaches. Our comparisons in Section 4 consider the CaMP score and the alternative p-values proposed by Getz et al. (2007), Forrest and Cavet (2007) and Rubin and Green (2007). These alternative statistics are used for implementing both the Benjamini and Hochberg (1995) method, the method discussed in Storey (2002) and the Empirical Bayes method.

4. Simulation study. In this section we compare alternative methods for identifying driver genes using the 10,000 simulation scenarios described in Section 2. The Wood et al. (2007) study considered both colon and breast tumors. Here we consider each tumor type separately, and repeat the same analysis.

Our simulation study compares the performance of alternative methods for ranking cancer genes and selecting putative cancer genes. Table 2 provides the average operating characteristics across all the simulated scenarios. All methods are set to control the FDR at 10% and 20% levels in turn. The empirical Bayes method estimates the FDR: the investigator controls the FDR by approximately matching the desired level α and the estimated proportion of false discoveries. Table 2 shows the average proportion of genes erroneously classified as drivers. Our results emphasize that operating characteristics are quite conservative when the FDR is controlled, on the basis of alternative p-values, using the Benjamini and Hochberg (1995) procedure. They also illustrate the importance of quantifying this conservative behavior by contrasting it with alternative methods such as the method proposed in Storey (2002) and the Empirical Bayes method. The Bayesian model of Section 2 allows us to simulate mutations across the genome, while the operating characteristics of the alternative methods shown in Table 2 provide

TABLE 2

Operating Characteristics of 9 *alternative procedures. Comparison between the Benjamini and Hochberg (BH), Storey (ST) and empirical Bayes (EB) methods. The average operating characteristics have been computed setting the FDR control at the 10% and 20% levels*

Method	Scores or p-values	False discoveries proportion		Average number of selected genes	
		$\alpha = 10\%$	$\alpha = 20\%$	$\alpha = 10\%$	$\alpha = 20\%$
		Colon-based simulations			
BH	CaMP score	0.101	0.218	150.4	221.8
BH	$p(\sum X^j_{mg} > x)$	0.074	0.146	115.1	198.7
BH	likelihood ratio	0.071	0.144	135.2	208.2
EB	CaMP score	0.106	0.232	162.3	242.6
EB	$p(\sum X^j_{mg} > x)$	0.100	0.238	147.8	250.4
EB	likelihood ratio	0.102	0.211	163.8	237.2
ST	CaMP score	0.104	0.220	157.3	235.4
ST	$p(\sum X^j_{mg} > x)$	0.099	0.232	146.5	242.4
ST	likelihood ratio	0.100	0.207	159.0	231.6
		Breast-based simulations			
BH	CaMP score	0.098	0.196	146.6	218.8
BH	$p(\sum X^j_{mg} > x)$	0.075	0.150	119.4	193.7
BH	likelihood ratio	0.074	0.141	131.7	205.3
EB	CaMP score	0.108	0.216	158.2	226.7
EB	$p(\sum X^j_{mg} > x)$	0.105	0.225	139.6	235.0
EB	likelihood ratio	0.098	0.209	153.5	223.4
ST	CaMP score	0.103	0.213	157.6	219.5
ST	$p(\sum X^j_{mg} > x)$	0.101	0.222	136.8	228.6
ST	likelihood ratio	0.096	0.207	147.6	221.2

approximations of their performances. The interpretation of the simulation study is anchored to the modeling assumptions formalized in Section 2. The results provide a solid basis for choosing among alternative methods.

All the methods considered produce lists of putative drivers genes with an average misclassification error below the 11% when the control of the FDR is set at the 10% level. If the desired α level is 10 %, our results indicate that, when the CaMP scores are adopted, with the Benjamini and Hochberg (1995) procedure, the average proportion of false discoveries is approximately equal to α, while under the empirical Bayes method and the Storey (2002) method, a slight excess of false discoveries is observed. We note, both in the colon and breast simulation studies, that the use of the likelihood ratios or the p-values $p(\sum X^j_{mg} > x)$ with the Benjamini and Hochberg (1995) procedure, on average, selects a substantially lower number of putative drivers than the alternative approaches. When the likelihood ratio and the p-value $p(\sum X^j_{mg} > x)$ are compared, under the empirical Bayes method and the Storey (2002) method, the likelihood ratio seams preferable; the average proportions of false discoveries are similar, but the likelihood ratio statistics select larger sets of putative drivers than the p-values $p(\sum X^j_{mg} > x)$. Also, when comparing the empirical Bayes method and the Storey (2002) method based on the likelihood ratio statistics, we observed only small variations both in the average proportion of false discoveries and in the average number of putative drivers. The operating characteristics when α is set at the 20% level confirm the conservative behavior of the Benjamini and Hochberg (1995) procedure with the likelihood ratios or the p-values $p(\sum X^j_{mg} > x)$, but also show departures from the target FDR α of the Empirical Bayes method and the Storey (2002) method. These results hold for both colorectal and breast cancer. When α is equal to 20%, the likelihood ratio seems preferable to the p-value $p(\sum X^j_{mg} > x)$, under both the empirical Bayes method and the Storey (2002) method; the likelihood ratios select putatitive drivers with an average misclassification error closer to the 20% target than the p-values.

The simulation-based comparison also allows us to asses the variability of the false discoveries proportion across scenarios, summarized in Figure 2. Each box plot is representative of the simulation-based distribution of the proportion of erroneously rejected hypothesis. The degree of variability is similar for all the methods considered; this similarity indicates that the average operating characteristics concisely reported in Table 2 are sufficient for a reliable evaluation of the methods.

To illustrate the importance of accounting for the two stages of the experimental design in computing tail probabilities $p(\sum X^j_{mg} > x)$, we repeated the analysis ignoring the two-stage structure. That is, we computed the tail probabilities under the erroneous assumption that the mutation counts were

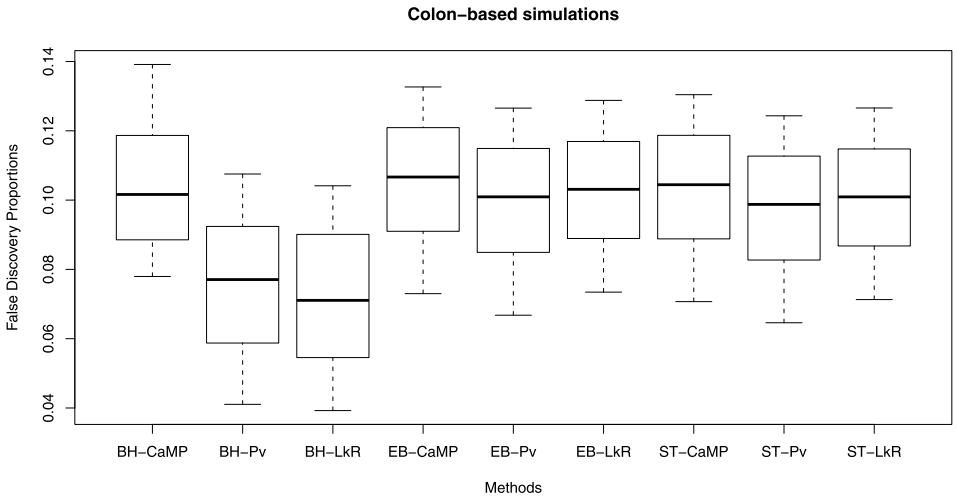

FIG. 2. *The distribution of the true false discovery proportions across simulated scenarios. Each box plot corresponds to one of the methods ($\alpha = 10\%$) considered in the simulation study. The lines of a box plot correspond to the median and to the* $10th$, $25th$, $75th$ *and* $90th$ *percentiles of the empirical distribution of the false discoveries proportion.*

observed in a single stage experiment. We observed a substantial reduction of the average number of selected genes, across simulation scenarios, when the tail probabilities are used for implementing the Benjamini and Hochberg (1995) procedure; with $\alpha = 20\%$ in the colon cancer and breast cancer cases, the averages become 165.6 and 157.0, respectively, and, with $\alpha = 10\%$, they decrease to 98.4 and 96.2.

Another important question concerns the fidelity of the ranking provided by these statistics. Figure 3 shows the Bayesian estimates of the left side of the ROC curves corresponding to each of the scores considered. The ROC curves are computed by separately estimating the scores' distributions across drivers and across passengers. The partial area under the curve is truncated at the 2% specificity level. This is equal to 2.05×10^{-3} for the log-likelihood ratio test statistic, 1.94×10^{-3} for the p-values $p(\sum X^j_{mg} > x)$ and 1.87×10^{-3} for the CaMP scores. The ROC curves estimates suggest that the log-likelihood ratio test statistic provides best discrimination, though differences are small.

5. Discussion. The investigation of somatic alterations is of primary interest in cancer research. Recent sequencing technologies have brought new insight into this question by revealing a landscape characterized by mutations involving a large number of driver genes, each altered in a relatively small fraction of tumors. When the earlier of these studies began to emerge, the cancer research community was faced with unexpected heterogeneity and

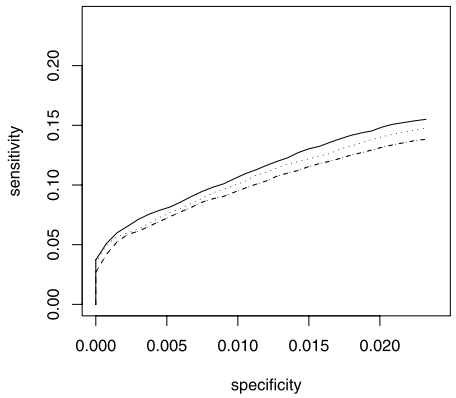

FIG. 3. *Estimated ROC curves for the log-likelihood ratio test statistics (solid line), the p-values* $p(\sum X_{mg}^j > x)$ *(dotted line) and the CaMP scores (dashed line).*

complexity, which required a significant change of perspective in both basic and clinical cancer investigations. The earliest genome-wide investigations [Sjöblom et al. (2006)] proposed this change of landscape based on a relative small number of samples. A key element in support of this proposal were estimates of the statistical significance of the reported list of driver genes [Sjöblom et al. (2006)]. These estimates were challenged: alternative approaches were proposed which would have led to reporting a drastically reduced number of candidate drivers at the same significance level.

Subsequent studies have provided strong supporting evidence for this new landscape, as well as validation for the driver role of many of the individual genes initially identified in Sjöblom et al. (2006). However, from a statistical standpoint, it remains very interesting to understand whether the initial conclusion was statistically sound based on the evidence available at the time, in part because similar problems will arise again in other cancer types and in other fields of genomics and evolutionary biology. Thus, our focus in this article has been the rigorous evaluation of methods for the identification of driver genes, with special emphasis on the methods that have been instrumental in the change of landscape we just described.

In the statistical literature, two common approaches for the evaluation of methodologies are asymptotic properties and scenario-based simulation studies. Asymptotic conclusions can be difficult to extrapolate to small samples. Scenario-based simulations can lack objectivity and comprehensiveness, and it can be a challenge to gauge whether the conclusions are applicable or not to a specific context of interest. To overcome these difficulties, we have proposed and implemented an alternative concept, which we hope will be very broadly applicable across statistics, and contribute to a more objective assessment of alternative methods. The idea is to construct a "super partes"

model that is (a) independent of the approaches being compared; (b) fits the data and available substantive knowledge well; and (c) can produce artificial data sets accounting for all relevant uncertainties, including parameter and potentially model uncertainty. This model is then used to simulate objective data-driven scenarios for method comparison. In this article our implementation of the super-partes model is based on Bayesian nonparametrics, an approach that can satisfy all three of the requirements above.

Strengths of the approach we proposed are the clear interpretation of both the assumptions captured by the Bayesian model and the average operating characteristics. The probability model's ability to reproduce the data structure allows to effectively interpret the results. The proposed evaluation scheme could be extended further to allow for more complex assumptions such as dependency among genes belonging to common functional pathways.

When applied to the controversy surrounding FDR control of early cancer genome studies, our method shows that the estimates provided in Sjöblom et al. (2006) are quite accurate despite the approximations used. Also, the Benjamini and Hochberg (1995) method is severely conservative. Last, the Empirical Bayes method and the Storey method based on likelihood ratios emerge as the preferred choices, though the margin of improvement is dependent on the control level.

Acknowledgments. We thank the Editor and two referees for helpful comments. We thank Ken Kinzler, Victor Velculescu and Bert Vogelstein for sharing their invaluable insight on the issues discussed in our analysis.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

The Cancer Genome Atlas project. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455** 1061–1068.

Cheng, C. and Pounds, S. (2007). False discovery rate paradigms for statistical analyses of microarray gene expression data. *Bioinformation* **1** 436–446.

Dudoit, S., Gilbert, H. and Laan, M. V. D. (2008). Resampling-based empirical Bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: Focus on the false discovery rate and simulation study. *Biom. J.* **50** 716–744.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** 71–103. MR1997066

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics. Camb. Ser. Stat. Probab. Math.* 223–273. Cambridge Univ. Press, Cambridge. MR2730665

EFRON, B. (2003). Robbins, empirical Bayes and microarrays. *Ann. Statist.* **31** 366–378. MR1983533

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

FARCOMENI, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* **17** 347–388. MR2526659

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FORREST, W. F. and CAVET, G. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers." *Science* **317** 1500 (author reply 1500).

GETZ, G., HÖFLING, H., MESIROV, J. P., GOLUB, T. R., MEYERSON, M., TIBSHIRANI, R. and LANDER, E. S. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers." *Science* **317** (5844) 1500.

GREENMAN, C., WOOSTER, R., FUTREAL, P. A., STRATTON, M. R. and EASTON, D. F. (2006). Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173** 2187–2198.

JONES, S., ZHANG, X., PARSONS, D. W., LIN, J. C. H., LEARY, R. J., ANGENENDT, P., MANKOO, P., CARTER, H., KAMIYAMA, H., JIMENO, A., HONG, S. M., FU, B., LIN, M. T., CALHOUN, E. S., KAMIYAMA, M., WALTER, K., NIKOLSKAYA, T., NIKOLSKY, Y., HARTIGAN, J., SMITH, D. R., HIDALGO, M., LEACH, S. D., KLEIN, A. P., JAFFEE, E. M., GOGGINS, M., MAITRA, A., IACOBUZIO-DONAHUE, C., ESHLEMAN, J. R., KERN, S. E., HRUBAN, R. H., KARCHIN, R., PAPADOPOULOS, N., PARMIGIANI, G., VOGELSTEIN, B., VELCULESCU, V. E. and KINZLER, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321** 1801–1806.

KRAFT, P. (2006). Efficient two-stage genome-wide association designs based on false positive report probabilities. *Pac. Symp. Biocomput.* 523–534.

MÜLLER, P. and QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19** 95–110. MR2082149

PARMIGIANI, G., LIN, J., BOCA, S. M., SJÖBLOM, T., JONES, S., WOOD, L. D., PARSONS, D. W., BARBER, T., BUCKHAULTS, P., MARKOWITZ, S. D., PARK, B. H., BACHMAN, K. E., PAPADOPOULOS, N., VOGELSTEIN, B., KINZLER, K. W. and VELCULESCU, V. E. (2007a). Response to comments on "The consensus coding sequences of human breast and colorectal cancers." *Science* **317** 1500.

PARMIGIANI, G., LIN, J., BOCA, S., SJÖBLOM, T., KINZLER, K., VELCULESCU, V. and VOGELSTEIN, B. (2007b). Statistical methods for the analysis of cancer genome sequencing. Working Paper 126. Johns Hopkins Univ., Dept. Biostatistics Working Papers. Available at http://www.bepress.com/jhubiostat/paper126.

PARMIGIANI, G., BOCA, S., LIN, J., KINZLER, K. and VELCULESCU, V. (2009). Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics* **93** 17–21.

PARSONS, D. W., JONES, S., ZHANG, X., LIN, J. C. H., LEARY, R. J., ANGENENDT, P., MANKOO, P., CARTER, H., SIU, I. M., GALLIA, G. L., OLIVI, A., MCLENDON, R., RASHEED, B. A., KEIR, S., NIKOLSKAYA, T., NIKOLSKY, Y., BUSAM, D. A., TEKLEAB, H., DIAZ, L. A., HARTIGAN, J., SMITH, D. R., STRAUSBERG, R. L., MARIE, S. K. N., SHINJO, S. M. O., YAN, H., RIGGINS, G. J., BIGNER, D. D., KARCHIN, R., PAPADOPOULOS, N., PARMIGIANI, G., VOGELSTEIN, B.,

Velculescu, V. E. and Kinzler, K. W. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science* **321** 1807–1812.

Rubin, A. F. and Green, P. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers." *Science* **317** 1500.

Satagopan, J. M. and Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genet. Epidemiol.* **25** 149–157.

Satagopan, J. M., Venkatraman, E. S. and Begg, C. B. (2004). Two-stage designs for gene-desease association studies with sample size constraints. *Biometrics* **60** 589–597. MR2089433

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E. and Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics* **58** 163–170. MR1891375

Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K. V., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W. and Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* **314** 268–274.

Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38** 209–213.

Smith, B. (2007). boa: An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software* **21** 1–37.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009). The cancer genome. *Nature* **458** 719–724.

Venturini, S., Dominici, F. and Parmigiani, G. (2008). Gamma shape mixtures for heavy-tailed distributions. *Ann. Appl. Statist.* **2** 756–776. MR2524355

Wang, H. and Stram, D. O. (2006). Optimal two-stage genome-wide association designs based on false discovery rate. *Comput. Statist. Data Anal.* **51** 457–465. MR2297463

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K. V., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V. K., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E. and Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science* **318** 1108–1113.

Dana-Farber Cancer Institute
3 Blackfan Circle, Boston
USA
and
Harvard School of Public Health
677 Huntington Avenue
Boston, Massachusetts 02115
USA
E-mail: ltrippa@jimmy.harvard.edu
gp@jimmy.harvard.edu